Cornell University
Department of Statistics and Data Science)
MPS Applied Statistics

Data Analysis Rural Households in Africa (CRS project)

Advisor:
Dr. David Matteson

Team member:
Xinrui Zhou(xz824), Katherine O'Connor(kso25), Yuhan Liu(yl3565), Xiaomeng Qin(xq72), Liuyuan Jiang(lj296), Chunyu Chen(cc2582)

January-May, 2023

Abstract

This project aims to analyze the food security indicators in Malawi using monthly household survey data collected by Catholic Relief Services from 2016 to 2023. Malawi faces consistent and recurring droughts, poverty, and stunted growth among children, making food security a crucial concern for the nation. The project seeks to determine household characteristics and behaviors that influence food security in rural areas over time, as well as the illnesses that are most closely associated with food insecurity. The project uses data dashboard for visualization and applies Random Forest, decision matrix, and Wilcoxon rank-sum test to analyze the data. The results of this analysis can be used to inform government policies and improve food poverty in Malawi.

1 Introduction

This data analysis project uses data provided by Catholic Relief Services. The data contains information from 2016 to 2022 in the region of Malawi. Local experts collect the questionnaires every month and record the answers in a system called CommCare. The data covers various aspects of the livelihoods of locals, including illness, education level, food supply, and dietary diversity. Additionally, the questionnaires also contain household information such as location, number of house members, and marital status. This project aims to analyze the changes in the livelihoods of locals over time, in order to provide suggestions for government support. Based on the background of the project, the following questions are of interest.

- What household characteristics or behaviors determine, or are most closely associated with, food security in rural areas over time?
- Which illnesses are most closely associated with food insecurity in rural areas of Malawi?

1.1 Literature review & background

Food security is a crucial aspect of livelihoods as it provides a reliable source of food for local individuals and households, which is vital for their survival and health. This is particularly important for households residing in

rural areas, where most people depend on agriculture for their livelihoods. Therefore, in our Malawi project, it is urgent to identify the factors that significantly impact local livelihoods and understand how these factors affect food security.

The Food Consumption Score (FCS) is used as one of the indicators to measure household food security. The article "Validation of the food programme's food consumption score and alternative indicators of household security" by Wiesmann et al. focuses on validating the FCS and other indicators of household food security. Their study involved a sample of 2000 households from rural areas of Ethiopia and Zimbabwe. The authors proved the validation of FCS as an indicator of household food security through quantitative and qualitative analysis. The study found that the FCS had high internal consistency and good test-retest reliability, indicating that the tool was consistent across different participants and over time. Additionally, the authors conducted in-depth interviews with households on their perceptions of the FCS to ensure that the FCS accurately reflects their experiences. The article also emphasized the importance of using the FCS along with other indicators, such as the Household Hunger Scale (HHS) and Coping Strategies Index (CSI), to fully explore various aspects of household food security.

The Household Hunger Scale (HHS) is an indicator that measures the severity of household hunger. The "Household Hunger Scale: Indicator Definition and Measurement Guide" written by Ballard et al. discusses the use and validation of the HHS as an indicator of household food security. The authors conducted extensive testing and validation across different countries to develop the HHS. They believed that the HHS is one of the significant measurements for researchers and governments concerning household food security.

The Coping Strategies Index (CSI) measures behaviors that people adopt when they cannot access enough food. It can be viewed as a quick and easy tool to administer and analyze and provide real-information to program managers. To construct the CSI for the Malawi area, we need to identify the locally relevant coping strategies in the study area, count the frequency of strategies, categorize and weight the strategies, and combine frequency and severity for analysis. As it can be used as a measure of the impact of food aid programs and an early warning indicator of an impending food crisis, it is a crucial indicator that requires analysis in this dataset.

The Household Dietary Diversity Score (HDDS) is defined as the number of different food groups consumed over a given reference period. A more

diversified diet is highly correlated with factors such as caloric and protein adequacy and household income, making it an effective indicator to measure food access. To construct the HDDS, the period of greatest food shortage will be chosen to collect data, and a series of yes/no questions will be asked to the person responsible for food preparation in the household.

Intuitively, household characteristics and macroeconomic status are associated with food security status. Walker (2021) found that those aged 35 to 49 and 50 to 64 were more likely to report food insecurity compared to the referent group aged 18 to 34 years. Therefore, age has a significant influence on food security. Although this article is written based on U.S. data, we believe that the result can be also applied to other countries, such as Malawi. Thus, an analysis of age is considered necessary.

According to Diamond-Smith (2019), food insecurity among married women in Nepal is associated with higher odds of experiencing intimate partner violence (IPV), specifically emotional and physical IPV. Therefore, a significant relationship exists between food insecurity and marital status. For instance, single women may be less likely to face family violence and have more food security. However, the opposite may also be true, as single women may lack protection in chaotic areas and experience more severe food security issues. Thus, investigating how marital status affects food security is a topic worth exploring.

Maharjan and Bauer (2017) examines the relationship between education and household food security in a rural area of Nepal using a cross-sectional survey of 240 households. The study measures household food security using the Household Food Insecurity Access Scale (HFIAS) and finds that individuals with higher education have greater knowledge and techniques related to food production and utilization, leading to increased food security. The study also reveals that educated individuals have greater knowledge of nutrition and food hygiene, contributing to a reduction in the prevalence of related health problems. The authors conclude that education level is highly related to household food security because individuals with higher education are more likely to contribute to local agriculture and help to increase the local economy. This article provides sufficient evidence that education level plays an essential role in food security in rural areas.

Gender is a potential factor that may significantly affect food security in rural areas of Malawi. The article "Gender and food security: An analysis of determinants of food security among smallholder farmers in Malawi" by Mhango et al. (2020) examines the role of gender in food security outcomes

among smallholder farmers in Malawi. The study employs a cross-sectional survey of 240 smallholder farmers and builds a logistic regression model to analyze the related data. The study finds that female farmers have lower levels of food security than male farmers, indicating that females in Malawi are not treated equally as males, and cultural beliefs often favor men over women. This situation results in female farmers having less chance of obtaining enough food and experiencing food insecurity for their families. Additionally, the study finds that female farmers have limited access to services such as education, land, and technology, which negatively affects their food security outcomes. The article highlights the significant role that gender plays in food security in the Malawi area, providing valuable insights for our project.

2 Data Description

2.1 Data Description and column selections

The RFMS database that is provided by Catholic Relief Services. The original database comprises 123,905 observations and 1,670 features. This dataset was obtained from a monthly survey that included a series of food-related questions. It contains basic information such as the case ID for different households, the survey round (seasonal data), survey month (monthly data), survey year (yearly data), and four main food security indicators: HHS, HDDS, FCS, and rCSI (as in Figure). These indicators measure the average food consumption, food diversity, and food-seeking strategies of the participants from Malawi rural households over the last month. To conduct data analysis, only features with more than 90% non-null values are considered. This leads to a cleaned RFMS dataset, which comprises 972 columns.

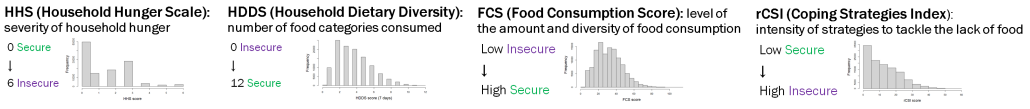


Figure 1: Histograms for the four indicators

There are 4 food security indicators that are of the most interest in the dataset as shown in Figure 1.

- HHS (Household Hunger Scale) is an indicator that describes the weighted sum of three severe food deprivation conditions faced by the participants over the last month, including a lack of food in the house, going to sleep hungry, and having no meals for the full day.

- HDDS (Household Dietary Diversity Score) generally shows the food diversity consumed by Malawi rural households in the past 7 days, including cereals, roots and tubers, vegetables, fruits, meat and poultry, eggs, fish and seafood, legumes and nuts, dairy, oil and fat, sugar, miscellaneous, and spices.

- FCS (Food Consumption Score) represents the sum of eight categories of foods consumed by Malawi rural households in the past 7 days, including staples, pulses, vegetables, fruits, meat and fish, milk, sugar, and oil and fat.

- rCSI (Reduced Coping Strategies Index) is measured by asking participants what kind of food strategies they made due to the food situation, and the weighted sum of five less severe strategies over the past week (borrowing, reducing quantity or frequency) is calculated from the data.

Other important indicators related to food security include roofing conditions, marriage status, education level, and age, among others. These indicators are all related to the food security indicators to some extent, and further analysis among them will be demonstrated in the data process part.

Due to the large number of columns, important columns (with more than 90% non-null values as defined before) are carefully analyzed through methods that will be discussed later. To conclude, the specific pools of features are selected to be analyzed for different questions.

Specifically, for categorical columns, different levels and situations are described in the data. For example, the variable 'survey_round' representing the different time points was selected for the time trend analysis to see the fluctuations of indicators. Marriage status, which includes several different statuses, such as 'head_monogamous', 'head_polygamous', 'head_separated', 'head_divorced', 'head_widowed', 'head_nevermarried', can be used to find the potential influence of different statuses on the food security indicators, such as plotting the histogram of HDDS grouping by marriage status. Education level, which is 'head_education_level,' can also be analyzed in the same way. In addition, other education-related indicators, such as head has

any formal education, head can read/write Chichewa, head can read/write English, and head currently enrolled in school, which are all binary variables, can be analyzed to learn about the influence of education further.

Other physical conditions may also have an effect on food security. For example, sufficient_housing, representing the average amount of housing needed, is divided into three levels: less than adequate, just adequate, and more than adequate, and it may reflect the economic situation of a household, which indicates the food intake in this household to some extent. Sufficient_clothing may have the same function as sufficient_housing. Furthermore, whether a household produces and stores its own food will be a potentially important factor in the amount of food intake.

For numerical columns, food consuming related numerical data, geographical related numeric data along with other numerical data are selected to later discover their significance in predicting the indicators, based on how much it contributes to reducing the randomness in the data (known as "gini impurity"). For example, Columns like 'days_item1'and 'days_item19' that represent the days the household consumed nsima and tomatoes within a week, 'FCS4_YN' and 'FCS8_YN' that represent any fruits ate and oil/fat ate within past 7 days. Those food related columns will be helpful because a diverse range of nutrient foods intake in rural areas could highly possibly play a crucial role in hunger scale and food security. Geographical data like 'gps_latitude' and 'gps_longtitude', which represent the latitude and longitude of the household location, are included in the analysis as well since both of them closely related to the climate and agriculture, and which may lead to indirect impact on the food intake as well as food security indicators.

Age related columns like 'members_5andunder', 'members_6to14', 'members_15to65' and 'members_over65' that represent the number of members within the corresponding age range in each household, are selected to analyze the impact of ages on the food security indicators. They are also used later to calculate the household age proportion like 'members_young.prop' and "members_elder.prop", which respectively describes the ratio of family members younger than 15 years old and the ratio older than 65 years old, to further analyze the impact of household age composition with Multinomial Logistic Ordinal Regression.

Sickness and illness related numerical columns are selected as well to discover the relationship between household health situation and all food security indicators. Columns 'man_sick' and 'woman_sick' are used to summarize and differentiate the sickness within the gender in the household, which re-

spectively represents the members of sick man and sick woman in the household. This will help us discover how the sickness situation within different gender could affect the food security indicators. Furthermore, columns related to sickness within different age groups like 'Sick0_2years', 'Sick3_5years', 'Sick6_15years', 'Sick16_65years' and 'SickOver65years' represent the number of sick individual within corresponding age range in each household. They are selected to further investigate the potential impact of the sickness situation among different age groups. Illness indicators are also selected to calculate the numerical 'illness score' in later visualization and analysis for each case ID by summing up the number of all illnesses reported, representing the number of illness types each individual suffered. This indicator will be used to visualize the individual illness behavior on food security indicators as well. It will be also included in later Time Series Analysis to discover patterns and trends of illness data.

Moreover, we also touch up a little with covid-related columns since it is obvious that this pandemic may have an impact on food supply, economics, education as well as health services, which may increase food insecurity specifically in rural areas. For example, columns like 'covidm42_Numbertest', which represent the number of members tested positive for the covid, and 'covid9a_maize' , which represents the maize available after covid. Our main objective is to find out what consequences of covid may impact the food security in malawi rural area.

Some other numerical columns are used as well based on the different analysis and modelings to further assist with better understanding of various impacts on food security indicators and to gather helpful information for society and government.

3 Methodologies

3.1 Decision Trees

A decision tree model is a popular machine learning algorithm used for classification and regression tasks. It is a tree-like model that uses a flowchart-like structure to map decisions and their possible consequences. At each node of the tree, a decision is made based on a feature, and the data is partitioned into subsets based on the outcome of that decision. This process is repeated recursively until a stopping criterion is met, such as reaching a maximum

tree depth or achieving a minimum number of samples at a leaf node.

One common approach to building decision trees is to use the Mean Decrease in Gini Impurity as a feature selection criterion. The Gini Impurity is a measure of how often a randomly chosen element from the set would be incorrectly labeled if it was randomly labeled according to the distribution of labels in the subset. The Mean Decrease in Gini Impurity is calculated by summing the reduction in impurity over all nodes that use a particular feature and averaging over all trees in an ensemble. This measure provides an estimate of the importance of each feature in the model, with higher values indicating greater importance.

By selecting features with high Mean Decrease in Gini Impurity, we can identify the most informative features for our decision tree model. This approach can improve the accuracy of our model by reducing the number of irrelevant or redundant features, while also making the model easier to interpret and more computationally efficient.

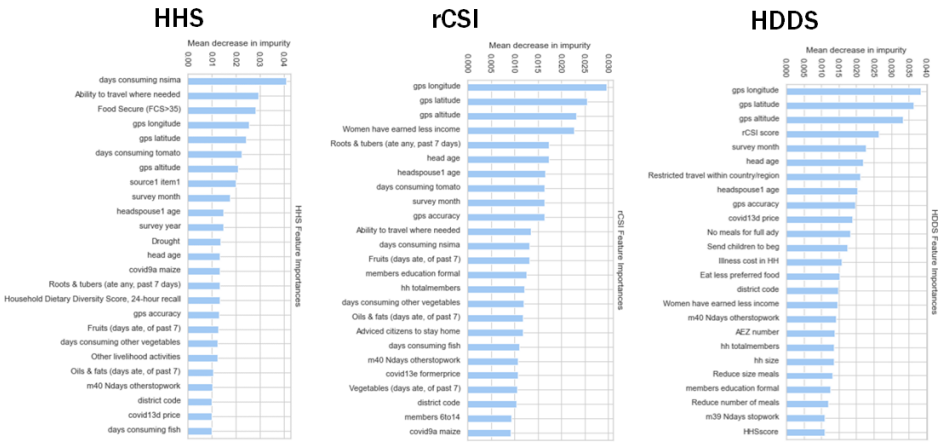


Figure 2: Features of the Greatest Mean Decrease in Gini Impurity

As presented in Figure 2, certain features are more strongly associated with food security than others. The top selected features (not only the ones that are presented in the figure) are further analyzed to explore their relationship with food security indicators. Specifically in this plot, the location associated features, time, and illness are of importance as well as the family characteristic features such as gender, marriage, education, and age that will be discussed in the following sessions.

3.2 Visualization Dashboard

Features that are of significance to food security indicators can be narrowed by looking into correlation coefficient and by looking into mean decrease in Gini Impurity. Given a selection of limited features, visualization are useful to look into the patterns. The dashboard consists of two main sections, one for categorical data and the other for numerical data.

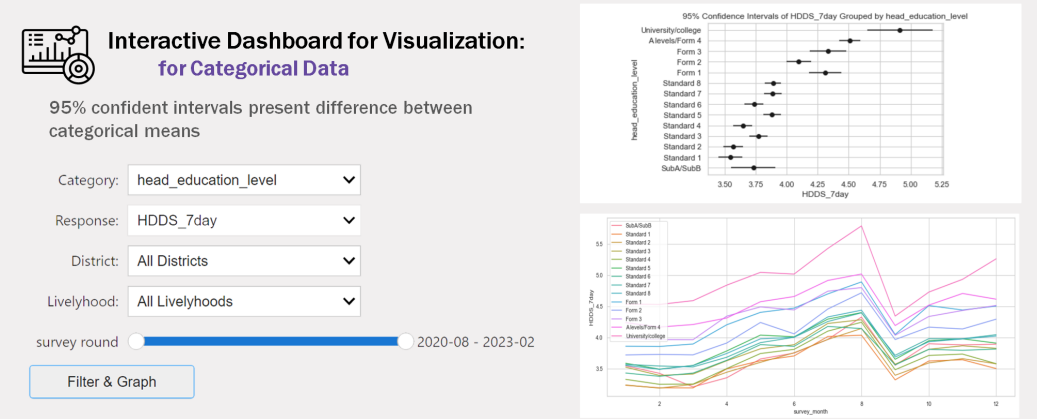


Figure 3: Visualization Dashboard for Categorical Data

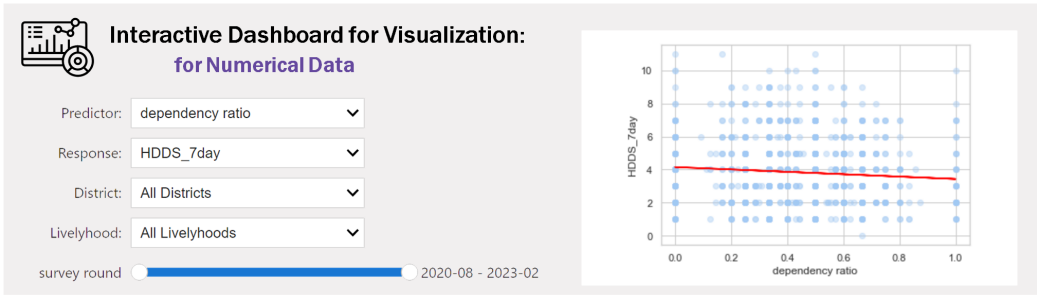
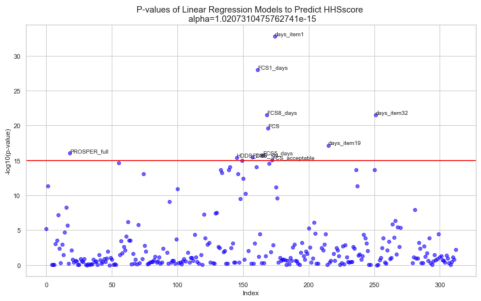


Figure 4: Visualization Dashboard for Numerical Data

For the categorical data section (Figure 3), the dashboard displays the average indicator scores for each survey month grouped by a common categorical variable. Users can filter the data by district, livelihood, and time to further refine the results. Additionally, the dashboard has implemented 95% Confidence Intervals to reflect the accuracy of the data based on the sample

size. With this feature, users can easily determine which categorical means are different from each other.

For the numerical data section (Figure 4), users can plot any indicator against any numerical variable of their choice. The dashboard generates a basic scatter plot and calculates a linear regression model from 1000 samples. It then conducts an F test on the beta coefficient of the linear model to determine the likelihood of it being non-zero. If the beta coefficient is non-zero, it implies a positive or negative association between the variable and indicator.



	Column	Documentation	p-value	r-value	slope	intercept
18	PROSPER_full	NaN	8.750867e-17	0.206	0.205	2.358
145	HDDS_24hr	Household Dietary Diversity Score, 24-hour recall	3.855836e-16	-0.202	-0.201	2.358
157	FCS5_YN	Meat & fish (any, past 7 days)	3.684382e-16	-0.202	-0.201	2.358
161	FCS1_days	Staples (days ate, of past 7)	1.110591e-28	-0.273	-0.272	2.358
165	FCS5_days	Meat & fish (days ate, of past 7)	1.822052e-16	-0.204	-0.203	2.358
168	FCS8_days	Oils & fats (days ate, of past 7)	3.073748e-22	-0.239	-0.238	2.358
169	FCS	Food Consumption Score (weighted, 7-day)	2.384048e-20	-0.228	-0.227	2.358
172	FCS_acceptable	Food Secure (FCS>35)	9.640037e-16	-0.199	-0.198	2.358
174	days_item1	(max) days_item1	1.641866e-33	-0.295	-0.294	2.358
215	days_item19	(max) days_item19	7.035552e-18	-0.213	-0.212	2.358
251	days_item32	(max) days_item32	3.073748e-22	-0.239	-0.238	2.358

Figure 5: Visualization Dashboard for Manhattan Plot

The dashboard additionally provides a feature to generate Manhattan plots (Figure 5), which are commonly used in quantitative genomics analysis. This feature repeats the process on all numeric columns from the entire dataset and graphs the p-values for each beta coefficient. With this, users can easily identify the most significant variables and their relationships to the indicator.

Moreover, the dashboard provides other visualization methods that are going to be discussed in the Result session.

3.3 Generalize Linear Regression & Multinomial Logistic Ordinal Regression

Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. It assumes a linear relationship between the variables, meaning that changes in the independent variables will result in a proportional change in the dependent variable.

The goal of linear regression is to find the best-fit line that describes the relationship between the variables, which can be used to predict the value of the dependent variable for a given set of independent variables. The line is defined by its intercept (the value of the dependent variable when all independent variables are zero) and slope (the change in the dependent variable for each unit increase in the independent variable). This method is one of the most commonly used one to explore the association between two numerical features.

Given the fact that the focused variables, namely HHS score, rCSI score, and HDDS score, are of the characteristics of ordinal categorical variables, Multinomial Logistic Ordinal Regression is applied to explore the effect of independent variables on them, therefore to answer the research questions.

Multinomial logistic ordinal regression is a statistical model that is used to analyze the relationship between a categorical dependent variable with two or more categories and one or more independent variables. The dependent variable is ordinal in nature, which means that it has a natural ordering, but the categories are not necessarily equidistant from each other.

The equation for multinomial logistic ordinal regression is as follows:

$$\log\left(\frac{p_{ij}}{p_{i(j-1)}}\right) = \alpha_j + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p \tag{1}$$

where:

- p_{ij} is the probability of observing an individual in category j given the value of the independent variables.
- $p_{i(j-1)}$ is the probability of observing an individual in category j-1 given the value of the independent variables.
- α_j is the intercept term for category j.
- $\beta_1, \beta_2, \ldots, \beta_p$ are the coefficients of the independent variables $x_1, x_2, \ldots, x_p$, respectively.

The methodology involves estimating the probabilities of each category of the dependent variable, given the values of the independent variables. The model is based on the assumption that the log-odds of being in a higher category versus a lower category are proportional across all levels of the independent variables. The equation expresses the log of the ratio of the

probability of being in category j versus category j-1 as a linear function of the independent variables. This can be transformed to obtain the probability of being in a particular category, given the values of the independent variables.

To estimate the model, maximum likelihood estimation is typically used. This involves finding the parameter estimates that maximize the likelihood of the observed data. The model assumes a certain distributional form for the errors, which may be specified as a normal, logistic, or other distribution.

4 Resutls

4.1 The Association of Household Characteristics or Behaviors with Food Security in Rural Areas over Time

4.1.1 Gender

To begin our analysis, we will examine the relationship between sex and three indicators: HHS score, rCSI score, and HDDS score. The plot below separates the data into two groups: females are represented by orange, and males are represented by blue. The first column of the plot uses each survey round as the x variable, while the second column uses the average of each survey month from January to December. Each row represents one of the three indicators.

In the first two rows of the plot, The average score for HHS score and rCSI score is consistently higher for females than for males. However, in the third row, the average HDDS score for females is always lower than that of males. A lower HHS or rCSI score indicates greater food insecurity, while a higher HDDS score indicates greater dietary diversity. Therefore, on average, females appear to have worse food consumption than males.

4.1.2 Marriage Status

Other than that, our team also discusses the relationship between marriage status and three indicators: HHS score, rCSI score, and HDDS score. The first plot in the left hand side below separates the data into two groups: divorced families are represented by orange, and non-divorced families are represented by blue. The second plot in the right hand side below separates the data into two groups: widows are represented by orange, and non-widows

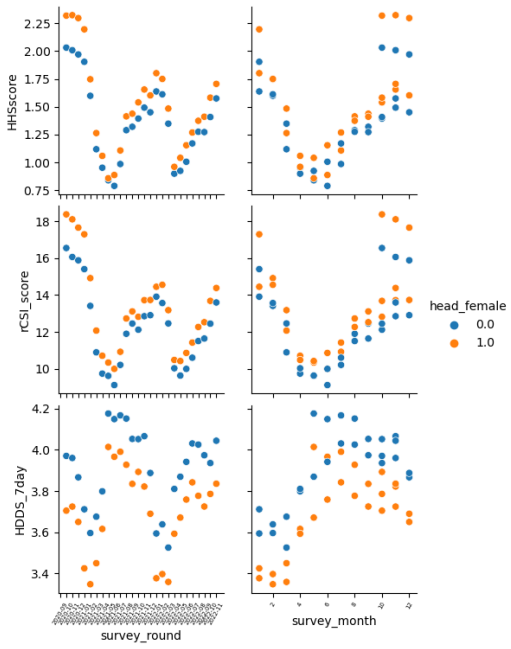


Figure 6: Three indicators V.S. Time in different Sex

are represented by blue. In both plots, the first column of the plot uses each survey round as the x variable, while the second column uses the average of each survey month from January to December. Each row represents one of the three indicators.

In the first two rows of the left hand side plot, the average HHS score is consistently higher for divorced families compared to non-divorced families. For the rCSI score, there is not much difference between the two groups. However, in the third row, the average HDDS score for divorced families is lower than that of undivorced families. Therefore, on average, divorced families appear to have worse dietary diversity than undivorced families. However, the difference is not significant enough to warrant additional support, as the rCSI scores are similar for both groups.

In the first two rows of the right hand side plot, The average score for HHS score and rCSI score is consistently higher for widows than for non-widows. However, in the third row, the average HDDS score for widows is always lower than that of non-widows. Therefore, on average, widows appear to have worse food consumption than non-widows.

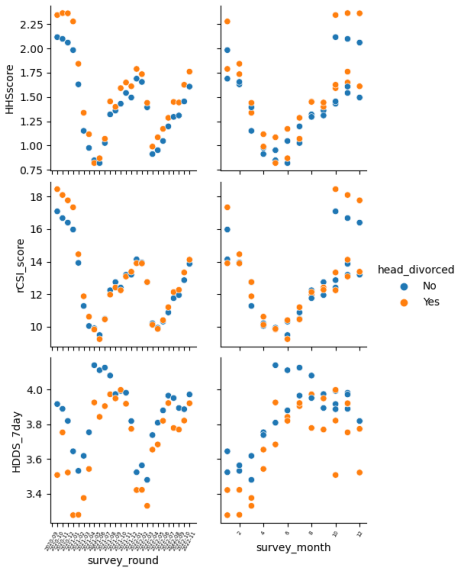


Figure 7: Three indicators V.S. Time in different family

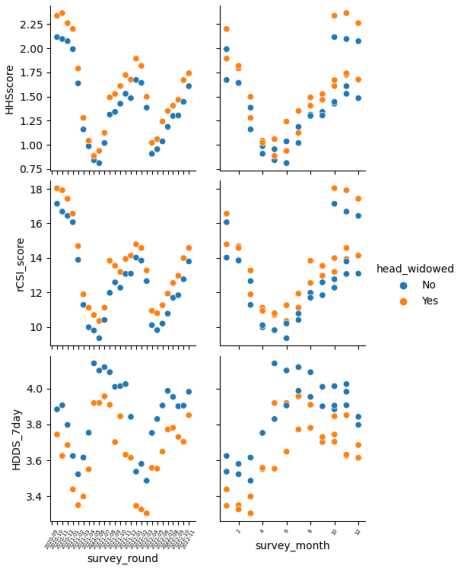


Figure 8: Three indicators V.S. Time in different female marriage status

4.1.3 Education Level

Moreover, educational levels of households may be important as an indicator of livelihood. The plot below separates the data into fourteen groups covering education level from kindergartens to university. The first column of the plot uses each survey round as the x variable, while the second column uses the average of each survey month from January to December. Each row represents one of the three indicators.

In the first two rows of the plot, the average HHS score and rCSI score are consistently higher for households with lower educational levels. However, in the third row, the average HDDS score is always lower for households with lower educational levels. Therefore, on average, households with lower educational levels appear to have worse dietary diversity and food consumption.

4.1.4 Age

To analyze the impact from family age proportion, features, namely 'members_5andunder', 'members_6to14', 'members_15to35', 'members_15to65', and

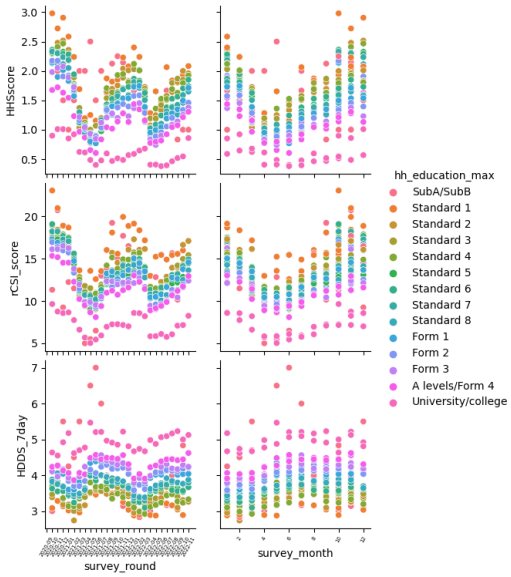


Figure 9: Three indicators V.S. Time in different Sex

'members_over65', are of interest. Considering the variability of the number of family members in different family, I analyze the proportion family members in specific age group and such proportion's quantitative impact on food security indicators.

- 'members_youth.prop': proportion of members aged ≤ 14.
- 'members_elder_grownup.prop': proportion of members aged $15 \leq 35$.
- 'members_young_grownup.prop': proportion of members aged $36 \leq 65$.
- 'members_elder.prop': proportion of members age ≥ 66.

By implement generalized linear regression model, we get the results in Table 1. Note that 'members_elder.prop' is not included in the model to avoid co-linearity.

Based on the output of the GLM model, we can see that the proportion of members in the young adult age group ('members_young_grownup.prop') and the proportion of members in the elder adult age group ('members_elder_grownup.prop') have a significant negative impact on the food security status of households.

Figure 10: Box Plot of Members Age $\leq$ 15 Proportion Grouped by HDDS.

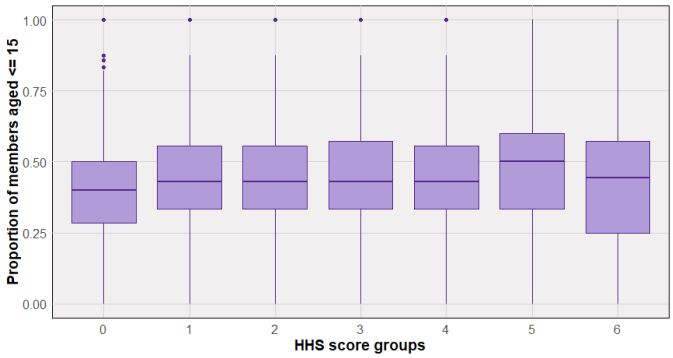


Figure 11: Box Plot of Members Age $\leq$ 15 Proportion Grouped by HHS.

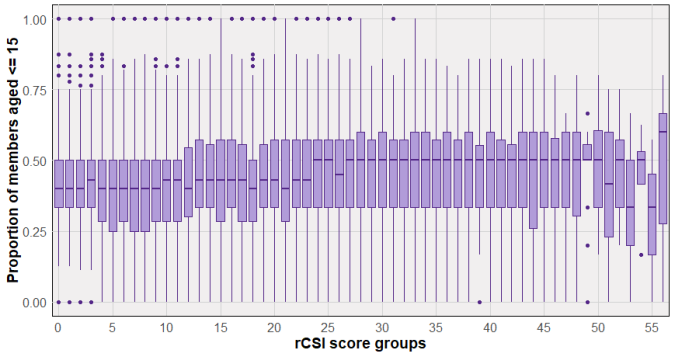


Figure 12: Box Plot of Members Age $\leq$ 15 Proportion Grouped by rCSI.

Coefficients	**Estimate**	**Std. Error**	**t value**	**Pr($>\|t\|$)**
(Intercept)	1.50723	0.03271	46.078	<2e-16 ***
members_youth.prop	0.08914	0.03829	2.328	0.0199 *
members_young_grownup.prop	-0.35534	0.03709	-9.581	<2e-16 ***
members_elder_grownup.prop	-0.12926	0.04141	-3.121	0.0018 **

Table 1: Summary of the linear model

This means that as the proportion of members in these age groups increases, the food security status of the household is likely to worsen.

On the other hand, the proportion of members in the youth age group ('members_youth.prop') has a significant positive impact on the food security status of households. This means that as the proportion of members in the youth age group increases, the food security status of the household is likely to improve.

The outcome (Table 2) from Multinomial Logistic Regression also corroborate the finding above. The proportion of kids member in family has negative impact on food security indicator HHS. Moreover, it can be concluded that the elder grownup (aged between 35-65) has minor contribution to the food security than younger grownup. This is consistent with the literature finding that is presented in the previous session.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept):1	-0.33652	0.04040	-8.329	< 2e-16 ***
(Intercept):2	0.18470	0.04039	4.573	4.81e-06 ***
(Intercept):3	0.87735	0.04053	21.648	< 2e-16 ***
(Intercept):4	2.79008	0.04289	65.057	< 2e-16 ***
(Intercept):5	3.46210	0.04546	76.152	< 2e-16 ***
(Intercept):6	4.00150	0.04899	81.679	< 2e-16 ***
members_youth.prop	0.21371	0.04717	-4.530	5.89e-06 ***
members_young_grownup.prop	-0.35362	0.04575	7.729	1.09e-14 ***
members_elder_grownup.prop	-0.09989	0.05103	1.957	0.0503 *

Table 2: Coefficients of the VGLM model

These findings suggest that policies and interventions aimed at improving food security should consider the age composition of households. Specifically, targeting households with a higher proportion of members in the young adult and elder adult age groups may be necessary to improve their food security

status. Meanwhile, households with a higher proportion of members in the youth age group may be more food secure and may require different interventions.

4.2 The association of illnesses with food security in rural areas of Malawi

4.2.1 Transition Matrix

The aim of this investigation was to identify the illnesses that are most closely associated with food insecurity in rural areas of Malawi. To achieve this objective, it is intuitive to calculate the the linear correlations between all illnesses and the food security indicators. Figure 13 shows the comparably important features. However, the coefficient is not big enough to conclude the significance. Therefore, transition matrices are applied to explore the illness's impact on the change of food security levels.

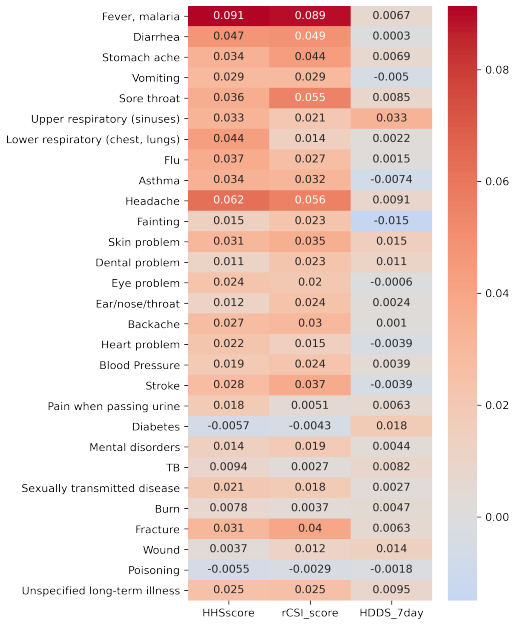


Figure 13: Correlation Coefficient Heat-map of Illness

The transition matrices represent the probabilities of families moving from one food security state to another over time. The rows and columns of

the matrix represent different states, and the numbers in the matrix represent the probability of transitioning from one state to another. The probabilities in each row must add up to one, as the system must always transition to some state. To demonstrate the impact of illness on food security status, we compare the transition matrix of families who remain healthy for two consecutive rounds of surveys with the matrices of families who are all healthy in the first round but have members who become ill in the following round.

Direct comparison can be done through looking at the heatmap. We can see that the probabilities of moving to worse food insecurity statuses are generally higher for families who are ill in phase 2 compared to those who are healthy in both phases. For example, if a family is in the Security status in phase 1, the probability of moving to Insecure status in phase 2 is 0.103 for the ill group, while it is only 0.069 for the healthy group. Similarly, the probability of moving to Catastrophe level status for the ill group is 0.520 when starting at Catastrophe level status in phase 1, compared to 0.596 for the healthy group.

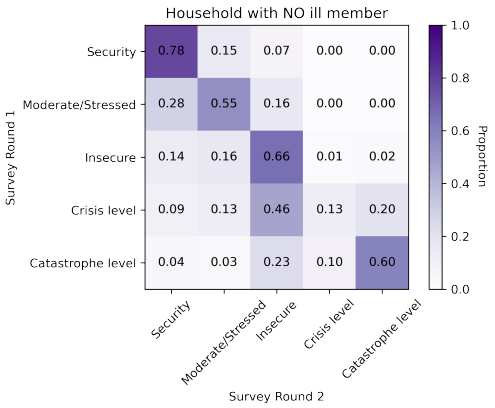


Figure 14: Healthy → Healthy

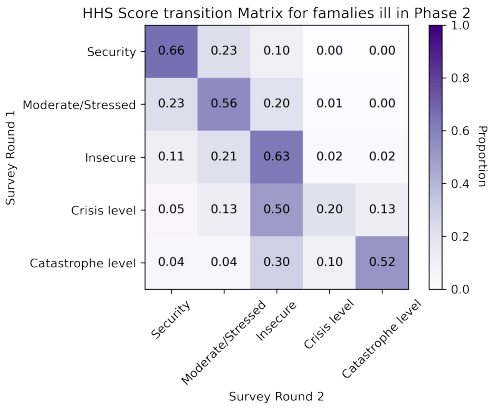


Figure 15: Healthy → Ill

Furthermore, statistics tests can be applied to quantify the difference. One-tailed Wilcoxon rank-sum test is applied to compare whether values in the left lower tail of the illness transition matrix is larger than that of all-time healthy families, which indicates the negative effect on food security status from illness. This test is appropriate here since the data are not normally distributed and the sample sizes are small with only left lower triangle values (e.g. 10 values for HHS score).

It can be seen from Table 4.2.1 that 16 diseases are significant by comparing to Figure 14. Note the diseases that has p-value equals 1 are the ones do not have enough observations to generate transition matrices. Moreover, Fever, malaria and headache are significant even by comparing to Figure 15.

4.2.2 Illness Score

Furthermore, a time series analysis is conducted on the illness dataset by creating a new column called "illness score" for each case ID by summing all illnesses reported over time. The time series plot (Figure 16) reveals that illness scores vary considerably over time, with some individuals experiencing relatively consistent levels of illness while others experience more pronounced fluctuations. Additionally, some seasonality is observed in the data, with higher illness scores during certain months of the year.

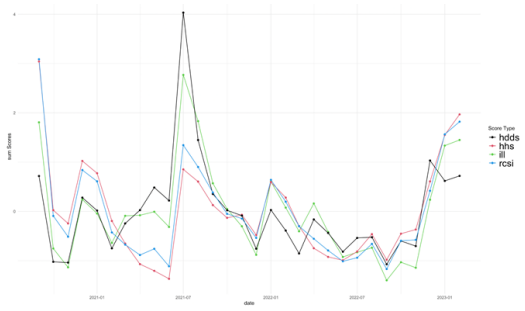


Figure 16: Time Series Plot of Illness Score and Food Security Indicators (Normalized)

The investigation also provides insight into the patterns and trends in the illness data by contextualizing the results of the correlation and GLM analyses. The overlap between illness and HDDS scores during the Covid period is attributed to the Malawi government's policy of distributing take-home rations for 600,000 learners, which may contribute to the higher HDDS scores. The findings of this investigation can inform the development of targeted interventions to address food insecurity and related health issues in rural areas of Malawi.

Variable	Illness	P-value
m36_illness_what12	Skin problem	0.002
m36_illness_what26	Fracture	0.002
m36_illness_what10	Headache	0.0049
m36_illness_what14	Eye problem	0.0049
m36_illness_what19	Stroke	0.0049
m36_illness_what16	Backache	0.0137
m36_illness_what17	Heart problem	0.0244
m36_illness_what3	Stomach ache	0.0322
m36_illness_what8	Flu	0.042
m36_illness_what9	Asthma	0.042
m36_illness_what20	Pain when passing urine	0.042
m36_illness_what29	Unspecified long-term illness	0.0527
m36_illness_what1	Fever, malaria	0.0654
m36_illness_what2	Diarrhea	0.0654
m36_illness_what7	Lower respiratory (chest, lungs)	0.0654
m36_illness_what6	Upper respiratory (sinuses)	0.0801
m36_illness_what15	Ear/nose/throat	0.1162
m36_illness_what5	Sore throat	0.1377
m36_illness_what13	Dental problem	0.1377
m36_illness_what4	Vomiting	0.1611
m36_illness_what27	Wound	0.1875
m36_illness_what18	Blood Pressure	0.2461
m36_illness_what11	Fainting	0.3125
m36_illness_what23	TB	0.3848
m36_illness_what25	Burn	0.5391
m36_illness_what21	Diabetes	1.0
m36_illness_what22	Mental disorders	1.0
m36_illness_what24	Sexually transmitted disease	1.0
m36_illness_what28	Poisoning	1.0

Table 3: Results of Wilcoxon signed-rank test

5 Conclusion

First, based on the scatterplots, some trends were found based on different indicators. For sex, the average scores for HHS and rCSI are consistently higher for females than for males, while the average HDDS score for females is always lower than for males. For divorced, the average HHS score is consistently higher for divorced households compared to non-divorced households, while the average HDDS score for divorced families is lower than that of undivorced families. For widows, the average score for HHS score and rCSI score is consistently higher for widows than for non-widows, while the average HDDS score for widows is always lower than that of non-widows. For educational level, the average HHS score and rCSI score are consistently higher for households with lower educational levels, while the average HDDS score is always lower for households with lower educational levels. Because a lower HHS or rCSI score indicates greater food insecurity, while a higher HDDS score indicates greater dietary diversity, a general conclusion can be drawn from the results. Overall, females, divorced, widowed, low-educated appear to have worse food consumption than males, non-divorced, non-widowed and high-educated households.

Our team created two dashboards to show the trends of indicators dynamically. The first dashboard shows average indicator scores by survey month of data grouped by a common categorical variable and can be filtered by district, livelihood and time. Confidence interval plot was drawn to see if different groups overlap with each other. The second dashboard is similar to the first one but focused on showing the association between association vectors and numerical variables. Manhattan plots were also used on all numeric columns to graph the p-values.

Furthermore, indicators of Age and Illness were explored by modeling. First, multinomial logistic ordinal regression was used for Age. Based on the result, households with a higher proportion of young or elderly members are at great risk of food insecurity. While illness was explored and investigated by GLM model and time series model. Based on the time series model, illness scores varied significantly over time, depending on individuals. Some individuals experienced relatively consistent levels of illness while others may experience more pronounced fluctuations. Seasonality also was found in the model. Moreover, transition matrices and Wilcoxon rank-sum test were applied to illness so that our team can explore the influence of illness on the change of food security levels.

For the modeling part, training and test data were divided to apply multiple models such as linear regression, polynomial regression, multilayer perceptrons, k-nearest neighbors, decision trees, random forests and so on. Based on the results, random forest would be the best model because of low testing MSE and the distribution of the predicted values was similar to distribution of the real ones. Ten random forests were constructed at different depths and we chose the most appropriate depth and graphed the importance of each feature with respect to predicting the indicator according to the mean decrease in gini impurity. Surprisingly, geographical indicators are the most important indicators to influence food security indicators, which means that while some findings appeared after our data analysis, much valuable information is hidden in the dataset, waiting for further research.

In conclusion, this report presents an analysis of the RFMS database provided by Catholic Relief Services, which comprises data from a monthly survey of food-related questions. Various visualization techniques and modeling were applied to explore the relationship between food security indicators and other indicators such as sex, marital status, and educational level. Random Forest was used to identify the most important features for further investigation, and geographical indicators would be the indicators that affect food security indicators most. A visualization dashboard was created by python functions to aid in data exploration and track fluctuations of indicators. Multinomial logistic ordinal regression, transition matrices, time series plots, and GLM models were applied to identify the association between age groups and food security as well as the association between illnesses and food security. The findings of this study can be used to inform policies aimed at improving food security and reducing poverty in Malawi.

References

[1] Wiesmann, D., Bassett, L., Benson, T. (2009). Validation of the food programme's food consumption score and alternative indicators of household food security. Washington, DC: International Food Policy Research Institute

[2] Ballard, T., Coates, J., Swindale, A., & Deitchler, M. (2011). Household Hunger Scale: Indicator Definition and Measurement Guide. Washington, D.C.: Food and Nutrition Technical Assistance III Project, FHI 360

[3] Daniel Maxwell, Richard Caldwell. (2008). The Coping Strategies Index, Field Methods Manual, Second Edition.

[4] Anne Swindale, Paula Bilinsky. (2006). Household Dietary Diversity Score (HDDS) for Measurement of Household Food Access: Indicator Guide

[5] Rebekah J. Walker, Emma Garacci, Aprill Z. Dawson, Joni S. Williams, Mukoso Ozieh, and Leonard E. Egede. (2021). Trends in Food Insecurity in the United States from 2011–2017: Disparities by Age, Sex, Race/Ethnicity, and Income

[6] Diamond-Smith N, Conroy AA, Tsai AC, Nekkanti M, Weiser SD.(2009). Food insecurity and intimate partner violence among married women in Nepal. J Glob Health

[7] Maharjan, M., Bauer, S. (2017). The effect of education on household food security in a rural area of Nepal. Food Security, 9(3), 529-542. doi: 10.1007/s12571-017-0661-9

[8] Mhango, W. G. (2020). Gender and food security: An analysis of determinants of food security among smallholder farmers in Malawi. Agriculture Food Security, 9(1), 1-14. doi: 10.1186/s40066-020-00261-5